And System For Data Analysis" (Serial No. 10/077,586), both of which are filed on even date herewith and incorporated herein in their entirety by reference.

[0077] Figure 14A is a GUI screen display depicting a table-like visualization of the data of Figure 13D according to an illustrative embodiment of the attribute reduction subsystem;

[0078] Figure 14B is a GUI screen display depicting the table-like visualization of Figure 14A subsequent to sorting according to an illustrative embodiment of the attribute reduction subsystem;

[0079] Figure 14C is a GUI screen display depicting a multiple line graph transformation of the data of Figure 14B according to an illustrative embodiment of the invention;

Delete paragraph [0080]

[0186] An extension of this technique is animating the display, where the attribute positions along the locus are consecutively shifted by a skip factor, such as one. That is, a fixed number of attributes (called a frame) are laid out on the locus. The number of attributes per frame is equal to the period of the time cycle data. The total number of attributes plotted typically includes several frame cycles worth. The radial visualization is then animated to show consecutive frames of data. Each individual display of the animation shows the same attributes, but with the attribute locations incremented by the skip factor. One advantage of this technique is that it can show data points that have unique time varying dependencies that are not seen in other visualizations. Some examples are discussed below with respect to Figures 13A-13D and 14A-14C.

[0191] Figure 14A is a GUI display screen 1307 of a table-like display 1344 of the type generated by the attribute reduction subsystem 102 of the invention. In the table 1344, the time samples T1-T100 are shown along the right margin. Each column of the table 1344 represents one of a thousand genes. The binned shading represents the gene expression values at each of the one hundred time samples T1-T100. However, with the time samples clustered and the records sorted by T1, in accord with the methods discussed herein, ten groups 1346a-1346k of

time intervals T1-T100 emerge. We can also see that the time samples T1, T11, T21, ..., T91 are in phase with each other, but ninety degrees out of phase with the time samples T6, T16, T26, ..., T96. Thus, the table 1344 provides additional information regarding analysis of time varying dependencies. The sinusoidal nature of the time dependencies of the data set of Figures 13A-13D and 14A-14B is further illustrated in the display 1311 of Figure 14C, which displays a multiple line graph representation of the data of Figure 14B. An illustrative process for such transformation is discussed above with respect to Figures 11A-11E.

[0192] As described above, according to the illustrative embodiment, the record categorization subsystem 104 employs the AP layout algorithm to determine the attribute positioning to realize the category separations of Figure 12C. Details of the illustrative AP algorithm are described next.

[0193] In one embodiment, a display screen of a radial visualization shows a 76 gene attribute subset of the Affymetrix™ gene set randomly arranged on the perimeter of a circular locus. The records (patients) are plotted within the locus in a manner such as described with respect to Figures 12A-12C. To test the 76 gene subset to determine if it is result-effective and/or to calibrate the radial visualization the illustrative record categorization subsystem 104 employs the AP algorithms.

[0194] The AP algorithms use class distinction metrics to assign the positions of the attributes on the locus. In the illustrative embodiment, the metric employed is t-statistics. The t-statistic is calculated for each column (gene attribute) by comparing all of the ALL values with all of the AML values in each column. The t-statistic is a standard statistical test for comparing two groups using the means and standard deviations. The t-statistic for each attribute determines the order of the attributes around the perimeter of the locus.

[0195] The genes or columns that have higher values for ALL are laid out in the top half of the locus, the genes or columns that have higher values for AML are laid out in the bottom half of the locus. The order of the genes are by t-statistic value. In the top half of the locus, the genes are ordered right to left with the most significant gene on the right and the least significant gene on the left. In the bottom half the genes are ordered with significance going from left to right.

[0196] The columns (genes) are laid out around the locus perimeter with the column that has the highest t-statistic (negative) value at the gene that has a higher mean for ALL than AML. One gene or column is most significant for having higher mean values for AML than ALL.

[0197] Use of the AP algorithms result in a relatively clean separation between the patients having AML-type leukemia and the patients having ALL-type leukemia.

[0198] Since the illustrative AP algorithm described above ranks the significance of the attributes, the operator may also employ the AP algorithms for attribute reduction. More specifically, subsets of the most significant attributes may be examined to determine further reduced, result-effective attribute subsets. By way of example, a radial visualization employing the top five most significant genes for ALL and AML shows that using this attribute subset, the AML-type patients and the ALL-type patients continue to clearly divide. Thus, the AP algorithms employed by illustrative record categorization subsystem 104 not only provide record categorization features, but also attribute reduction features.

[0228] Figure 24 is a GUI screen image 2400 depicting a multi-dimensional polygonal visualization 2402 in the display panel 1706 according to an illustrative embodiment of the invention. The polygonal visualization 2402 includes a number of records 2408 disposed at locations determined in relation to a plurality of attributes 2404 by way of the methodology discussed above. The attributes on the locus of the circle are extended to form lines. This line now represents the attribute with the minimum value at one end of the line and the maximum value at the other. Thus, it is an axis and this yields a polygonal display. Each record in the display has a value for that attribute and the line from the attribute value points to the record. In many cases the values for each attribute have a distribution which is represented on the attribute line, thus yielding multiple lines pointing to the records. This is similar to parallel coordinates for which the lines represent the axes. In Figure 24, the control panel 1708 includes a button 2404 which enables an operator to activate global parameters during the display of the polygonal visualization 2402, and slider controls 2410a and 2410b which control the resolution of data in the X and Y directions, respectively. The control panel 1708 of Figure 24 also includes a plurality of check boxes 2412a-2412f. The check boxes 2412a-2412f control whether a floating

probe is displayed, and if so, the features of the information displayed using to the floating probe. The floating point probe displays actual attribute values. The control panel 1708 further includes a pull-down menu 2414 which selects a region mode. The region mode menu 2414 enables an operator to select a region of the visualization 2402 for display and/or analysis by way of a pointing device, such as a mouse. The control panel 1708 also provides a series of user interactive dialog boxes 2416a-2416e for manipulating the forces applied to the records 2808 during plotting on the locus 2406. An operator enters a desired force equation in the dialog box 2418. To enter a force equation into any of the dialog boxes 2416a-2416e, the operator enters the force equation in the dialog box 2418 and then selects one or more dialog boxes 2416a-2416e to indicate to which of the attributes the entered force equation is to be applied.

[0238] Figure 32 is a GUI screen image 3200 illustrating the "Data" pull-down menu 1716f. In Figure 32, the "Data" 1716f menu has been activated. The resulting pull-down menu 3200 includes entries "Do All Sort" 3202a, "Sum Sort of Records" 3202b, "Show Table ..." 3202c, "Set Missing Values" 3202d, and "Pivot Data" 3202e. The use of the pull-down menu commands 3202a-3202e follows the conventional method of highlighting a command with a mouse or other pointing device and activating the highlighted command with an action such as a mouse button click. The "Sort" commands 3202a and 3202b are used to sort by rows or columns in the gray scale binned table 3204. The "Show Table ..." command 5360c displays the numerical data corresponding to entries in the table 3204. The "Set Missing Values" command 3202d inserts missing values according to the condition assigned by the operator for missing values, as discussed above, or in the absence of such action by the operator, inserting default values. The "Pivot Data" command 3202e causes the exchange of rows and columns in the table 3204.

[0279] During practice of the invention, it has been discovered that various other subgroups of the 6817 genes, the expression products of which were tested in Golub *et al.* (1999) *supra*, can be used to identify and distinguish individuals with AML, B ALL and T ALL. Three classes of genes comprising 76 genes, 57 genes and 3 genes were identified using different forms of the algorithms described herein. For example, 76 gene products have been identified using the methods and systems described herein which can be used to identify AML patients that respond differently to treatment regimes (see, Figure 34). Figure 34 shows criteria for distinguishing

between individuals 3402 with AML that respond to chemotherapy from those 3404 that do not respond to chemotherapy. The 76 genes are identified in Table 1 below together with their GenBank accession numbers, the sequences of which are incorporated herein by reference. The sequences can be obtained through the National Center for Biotechnology Information (NCBI) web site at www.ncbi.nlm.nih.gov.

[0301] Figure 49 depicts a GUI screen image 4900 of parameters for the AP algorithm, as described above. The GUI screen image 4900 shows a "Set Discrimination Threshold" dialog box 4902 that enables the selection of parameters for class distinction. The "Set Discrimination Threshold" dialog box 4902 enables the selection of GS, option 1, and option 2, and the selection of a positive differential selection or a negative differential selection. The GS, option 1, and option 2 select differential statistical measures for laying out the attributes. Further, a significance level is employed upon the selection of the "Use Significance Level" checkbox 4904. Moreover, the dialog box 4902 enables an input of a threshold value 4906 and/or a maximum class size 4908.